COURSERA FINAL PROJECT

SPECIALIZED MODELS: TIME SERIES AND SURVIVAL ANALYSIS

Time Series Forecasting to predict grocery sales at Favorita stores

Project presented by

Raziel Amador Rios

to obtain the Coursera Certificate

Main object: produce a reliable forecasting based on sales time-series data from a retail store (Favorita stores, an Ecuadorian-based company).

Barcelona, January 2023

Table of Contents

Time Series Forecasting

I

Main objective

The **main object** of the project is to generate a reliable **time series forecasting** based on sales from Favorita stores, which is a grocery retail store based in Ecuador. Producing an accurate forecast could lead to decreased food waste related to overstocking and improve customer satisfaction.

All the code can be found in the following GitHub repository, further information can be found in Supplementary information.

II

Data description

The dataset comes from **Kaggle**, named: Store Sales - Time Series Forecasting. The Kaggle dataset contains sales data from *Corporación Favorita*, a large Ecuadorian-based grocery retailer. Kaggle provides you with 7 diffeent files (see Table 1 for more details).

Number	File Name	Description
1	holiday_events.csv	Relevant holidays in Ecuador
2	oil.csv	Oil prices from 2013 to 2017
3	sample_submission.csv	submission example
4	stores.csv	Stores metadata
5	test.csv	Stores and family products
6	train.csv	Sales by store and product-family from 2013-01 to 2017-08
7	transactions.csv	Number of transactions by store

Table 1: Kaggle files description. Files are alphabetically sorted.

The training data represents 99% of the data, including dates from 2013-01-01 to 2017-08-16 (55.5 months), 54 stores placed in different cities within Ecuador, and 33 family-products (see Figure 1). The testing data includes dates from 2017-08-16 to 2017-08-31 (15 days).

Figure 1: Summary of the training dataset. The cluster information denotes similarity between stores.

III

Data exploration and data cleaning

Performing an exploratory data analysis (eda) of the sales from the training dataset, we can observe that grocery I, beverages, and produce are the top 3 most consumed products (see Figure 2) . Additionally, store type A, and cluster 5 are the most frequent among their classification.

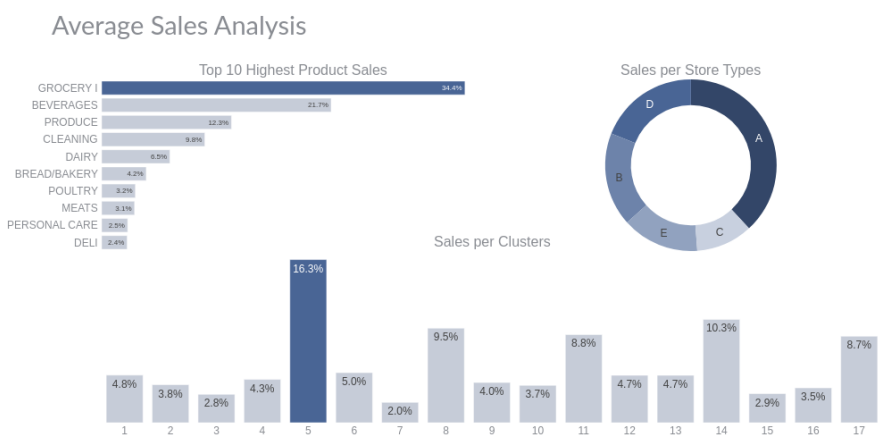

Figure 2: EDA of sales. The plot describes the sales by product, store type, and per cluster (left, right, and below, respectively). Darker blue represents higher sales.

III.1. Stationarity and seasonality analyses

The sales represented by Figure 3 shows low-peaks at the end of each year, which is explained because the stores are closed at new years. Moreover, we can observe a pattern at each year, with increased sales at the end of the year which overlaps with the Christmas eves, suggesting a seasonal pattern. These patterns are highlighted

after smoothing the sales data with a 7 days moving average (the continuous red lines from the Figure 3).

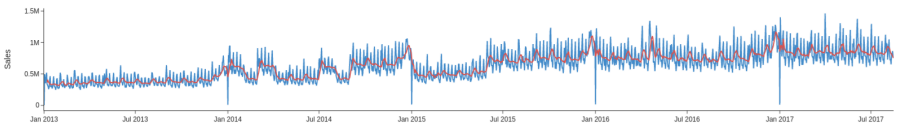

Figure 3: Target time series data. The y-axis represents the training sales from 2013-01-01 to 2017-08-15, and the x-axis shows the time in days (1,087 days). The sales were aggregated by all stores and products (see Figure 1). The raw data, and smoothed data (7 days moving average) are denoted by the continuous blue, and red lines, respectively.

The stationary and seasonality are key components to select the adequate machine learning model (such as ARIMA, SARIMA, etc.) to generate reliable forecasting. A stationary time series is defined, if their statistical properties such as mean, and variance are all constant, and independent of time. In consequence, we implemented a *Dickey-Fuller test* to assess stationarity. We obtained a *p-value* of 0.09, using a significance level of 0.05, we can reject the null hypothesis (the series is stationary) concluding that our series is non-stationarity. Obtaining valuable information about which models are more suitable.

III.2. Time series Decomposition

Appendix

IV

Supplementary information

This work was written with emacs[1] using LaTeX[2], using only **Free and Open Source software**. All the computational analysis were carried out using Linux-based distributions. The figures were generated with Python (matplotlib[3]/seaborn[4]/plotly[5]) and Inkscape[6].

Contact: razielar@gmail.com

[1] https://www.gnu.org/software/emacs/
[2] https://www.latex-project.org/
[3] https://matplotlib.org/
[4] https://seaborn.pydata.org/
[5] https://plotly.com/
[6] https://inkscape.org/